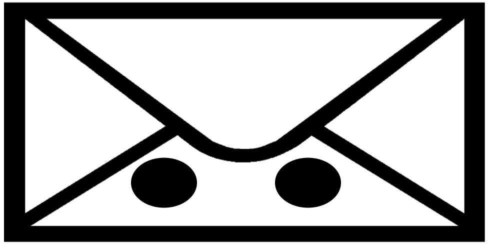
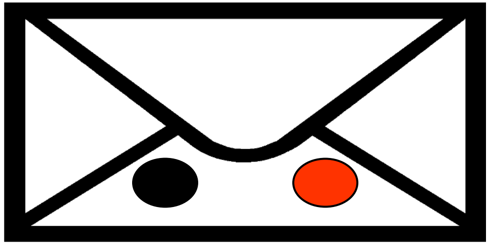
Two Envelopes Problem

- We have two envelopes:
  - E_1 has two black balls, E_2 has one black, one red
  - The **red** one is worth \$100. Others, zero
  - Open an envelope, see one ball. Then, can switch (or not).
  - You see a black ball. **Switch?**





Two Envelopes Solution

- Let's solve it.

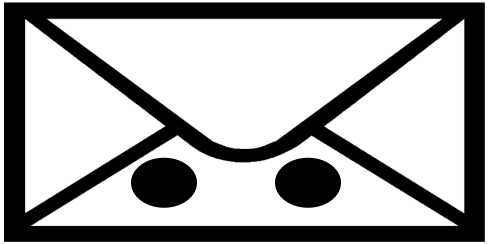
$$P(E_1|\text{Black ball}) = \frac{P(\text{Black ball}|E_1)P(E_1)}{P(\text{Black ball})}$$

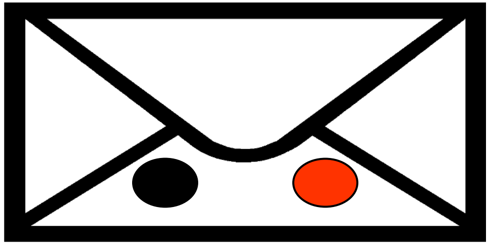
- Now plug in:

$$P(E_1|\text{Black ball}) = \frac{1 \times \frac{1}{2}}{P(\text{Black ball})}$$

$$P(E_2|\text{Black ball}) = \frac{\frac{1}{2} \times \frac{1}{2}}{P(\text{Black ball})}$$

So switch!





Break & Quiz

Q 1.1: Suppose P is false, Q is true, and R is true. Does this assignment satisfy

(i) $\neg(\neg p \rightarrow \neg q) \land r$

(ii) $(\neg p \lor \neg q) \rightarrow (p \lor \neg r)$

- A. Both
- B. Neither
- C. Just (i)
- D. Just (ii)

Break & Quiz

Q 1.1: Suppose P is false, Q is true, and R is true. Does this assignment satisfy

(i) $\neg(\neg p \rightarrow \neg q) \wedge r$

(ii) $(\neg p \vee \neg q) \rightarrow (p \vee \neg r)$

- A. Both
- B. Neither
- **C. Just (i)**
- D. Just (ii)

Break & Quiz

Q 1.2: Let A =“Aldo is Italian” and B =“Bob is English”. Formalize “Aldo is Italian or if Aldo isn’t Italian then Bob is English”.

- a. $A \lor (\neg A \rightarrow B)$
- b. $A \lor B$
- c. $A \lor (A \rightarrow B)$
- d. $A \rightarrow B$

Break & Quiz

Q 1.2: Let A =“Aldo is Italian” and B =“Bob is English”. Formalize “Aldo is Italian or if Aldo isn’t Italian then Bob is English”.

- **a. $A \vee (\neg A \rightarrow B)$**
- **b. $A \vee B$ (equivalent!)**
- c. $A \vee (A \rightarrow B)$
- d. $A \rightarrow B$

Break & Quiz

Q 1.3: How many different assignments can there be to $(x_1 \wedge y_1) \vee (x_2 \wedge y_2) \vee \ldots \vee (x_n \wedge y_n)$

- A. 2
- B. 2^n
- C. 2^{2n}
- D. 2n

Break & Quiz

Q 1.3: How many different assignments can there be to $(x_1 \wedge y_1) \vee (x_2 \wedge y_2) \vee \ldots \vee (x_n \wedge y_n)$

- A. 2
- B. 2^n
- **C. 2^{2n}**
- D. 2n

Break & Quiz

Q 2.1: Which has more rows: a truth table on *n* symbols, or a joint distribution table on *n* binary random variables?

- A. Truth table
- B. Distribution
- C. Same size
- D. It depends

Break & Quiz

Q 2.1: Which has more rows: a truth table on *n* symbols, or a joint distribution table on *n* binary random variables?

- A. Truth table
- B. Distribution
- **C. Same size**
- D. It depends

Break & Quiz

Q 1.1: Which of the below are bigrams from the sentence "It is cold outside today".

- A. It is
- B. cold today
- C. is cold
- D. A & C

Break & Quiz

Q 1.1: Which of the below are bigrams from the sentence "It is cold outside today".

- A. It is
- B. cold today
- C. is cold
- **D. A & C**

Break & Quiz

Q 1.2: Smoothing is increasingly useful for n-grams when

- A. n gets larger
- B. n gets smaller
- C. always the same
- D. n larger than 10

Break & Quiz

Q 1.2: Smoothing is increasingly useful for n-grams when

- **A. n gets larger**
- B. n gets smaller
- C. always the same
- D. n larger than 10

Break & Quiz

Q 2.1: What is the perplexity for a sequence of *n* digits 0-9? All occur with equal probability.

$$\mathrm{PP}(W) = P(w_1, w_2, \ldots, w_n)^{-\frac{1}{n}}$$

- A. 10
- B. 1/10
- C. 10^n
- D. 0

Break & Quiz

Q 2.1: What is the perplexity for a sequence of *n* digits 0-9? All occur with equal probability.

$$\mathrm{PP}(W) = P(w_1, w_2, \ldots, w_n)^{-\frac{1}{n}}$$

- **A. 10**
- B. 1/10
- C. 10^n
- D. 0

Quiz Break

Q1-1: Which is true about feature vectors?

A. Feature vectors can have at most 10 dimensions
B. Feature vectors have only numeric values
C. The raw image can also be used as the feature vector
D. Text data don't have feature vectors

A. Feature vectors can be in high dimen.
B. Some feature vectors can have other types of values like strings
D. Bag-of-words is a type of feature vector for text

Quiz Break

Q1-2: Which of the following is not a common task of supervised learning?

A. Object detection (predicting bounding box from raw images)
B. Classification
C. Regression
D. Dimensionality reduction

Quiz Break

Q1-2: Which of the following is not a common task of supervised learning?

A. Object detection (predicting bounding box from raw images)
B. Classification
C. Regression
D. Dimensionality reduction

Quiz Break

Q2-1: Which is true about machine learning?

A. The process doesn’t involve human inputs
B. The machine is given the training and test data for learning
C. In clustering, the training data also have labels for learning
D. Supervised learning involves labeled data

Quiz Break

Q2-1: Which is true about machine learning?

A. The process doesn't involve human inputs
B. The machine is given the training and test data for learning
C. In clustering, the training data also have labels for learning
D. Supervised learning involves labeled data

A. The labels are human inputs
B. The machine should not have test data for learning
C. No labels available for clustering

Quiz Break

Q2-2: Which is true about unsupervised learning?

A. There are only 2 unsupervised learning algorithms
B. Kmeans clustering is a type of hierarchical clustering
C. Kmeans algorithm automatically determines the number of clusters k
D. Unsupervised learning is widely used in many applications

Quiz Break

Q2-2: Which is true about unsupervised learning?

A. There are only 2 unsupervised learning algorithms
B. Kmeans clustering is a type of hierarchical clustering
C. Kmeans algorithm automatically determines the number of clusters k
D. Unsupervised learning is widely used in many applications

Break & Quiz

Q 1.1: We have two datasets: a social network dataset S_1 which shows which individuals are friends with each other along with image dataset $S_{2.}$

What kind of clustering can we do? Assume we do not make additional data transformations.

- A. k-means on both S_1 and S_2
- B. graph-based on S_1 and k-means on S_2
- C. k-means on S_1 and graph-based on S_2
- D. hierarchical on S_1 and graph-based on S_2

Break & Quiz

Q 1.1: We have two datasets: a social network dataset S_1 which shows which individuals are friends with each other along with image dataset $S_{2.}$

What kind of clustering can we do? Assume we do not make additional data transformations.

- A. k-means on both S_1 and S_2
- **B. graph-based on S_1 and k-means on S_2**
- C. k-means on S_1 and graph-based on S_2
- D. hierarchical on S_1 and graph-based on S_2

Break & Quiz

Q 1.1: We have two datasets: a social network dataset S_1 which shows which individuals are friends with each other along with image dataset $S_{2.}$
What kind of clustering can we do? Assume we do not make additional data transformations.

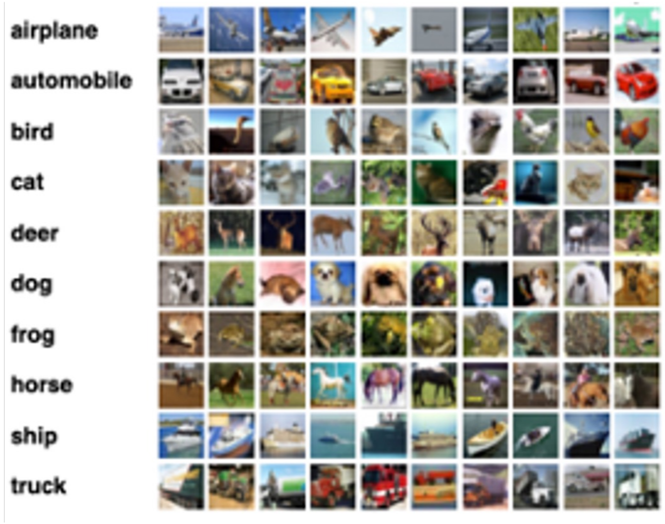
- A. k-means on both S_1 and S_2 **(No: can't do k-means on graph)**
- **B. graph-based on S_1 and k-means on S_2**
- C. k-means on S_1 and graph-based on S **(Same as A)**
- D. hierarchical on S_1 and graph-based on S_2 **(No: S_2 is not a graph)**

Break & Quiz

Q 1.2: The CIFAR-10 dataset contains 32x32 images labeled with one of 10 classes. What could we use it for?

(i) Supervised learning (ii) PCA (iii) k-means clustering

- A. Only (i)
- B. Only (ii) and (iii)
- C. Only (i) and (ii)
- D. All of them



Break & Quiz

Q 1.2: The CIFAR-10 dataset contains 32x32 images labeled with one of 10 classes. What could we use it for?

(i) Supervised learning (ii) PCA (iii) k-means clustering

- A. Only (i)
- B. Only (ii) and (iii)
- C. Only (i) and (ii)
- **D. All of them**

Break & Quiz

Q 1.2: The CIFAR-10 dataset contains 32x32 images labeled with one of 10 classes. What could we use it for?

(i) Supervised learning (ii) PCA (iii) k-means clustering

- (i) **Yes: train an image classifier; have labels)**
- (ii) **Yes: run PCA on image vectors to reduce dimensionality**
- (iii) **Yes: can cluster image vectors with k-means**
- **D. All of them**

Break & Quiz

Q 2.1: Can we do t-SNE on NLP (words) or graph datasets?

- A. Never
- B. Yes, after running PCA on them
- C. Yes, after mapping them into R^{d} (ie, embedding)
- D. Yes, after running hierarchical clustering on them

Break & Quiz

Q 2.1: Can we do t-SNE on NLP (words) or graph datasets?

- A. Never
- B. Yes, after running PCA on them
- **C. Yes, after mapping them into R^d (ie, embedding)**
- D. Yes, after running hierarchical clustering on them

Break & Quiz

Q 2.1: Can we do t-SNE on NLP (words) or graph datasets?

- A. Never **(No: too strong)**
- B. Yes, after running PCA on them **(No: can't run PCA on words or graphs directly. Need vectors)**
- **C. Yes, after mapping them into R^d (ie, embedding)**
- D. Yes, after running hierarchical clustering on them **(No: hierarchical clustering gives us a graph)**

Break & Quiz

Q 2.1: When we train a model, we are

- A. Optimizing the parameters and keeping the features fixed.
- B. Optimizing the features and keeping the parameters fixed.
- C. Optimizing the parameters and the features.
- D. Keeping parameters and features fixed and changing the predictions.

Break & Quiz

Q 2.1: When we train a model, we are

- **A. Optimizing the parameters and keeping the features fixed.**
- B. Optimizing the features and keeping the parameters fixed.
- C. Optimizing the parameters and the features.
- D. Keeping parameters and features fixed and changing the predictions.

Break & Quiz

Q 2.1: When we train a model, we are

- **A. Optimizing the parameters and keeping the features fixed.**
- B. Optimizing the features and keeping the parameters fixed) (Feature vectors *x*i don’t change during training).
- C. Optimizing the parameters and the features. (Same as B)
- D. Keeping parameters and features fixed and changing the predictions. (We can’t train if we don’t change the parameters)

Break & Quiz

- **Q 2.2**: You have trained a classifier, and you find there is significantly **higher** loss on the test set than the training set. What is likely the case?

- A. You have accidentally trained your classifier on the test set.
- B. Your classifier is generalizing well.
- C. Your classifier is generalizing poorly.
- D. Your classifier is ready for use.

Break & Quiz

- **Q 2.2**: You have trained a classifier, and you find there is significantly **higher** loss on the test set than the training set. What is likely the case?

- A. You have accidentally trained your classifier on the test set.
- B. Your classifier is generalizing well.
- **C. Your classifier is generalizing poorly.**
- D. Your classifier is ready for use.

Break & Quiz

- **Q 2.2**: You have trained a classifier, and you find there is significantly **higher** loss on the test set than the training set. What is likely the case?

- A. You have accidentally trained your classifier on the test set. **(No, this would make test loss lower)**
- B. Your classifier is generalizing well. **(No, test loss is high means poor generalization)**
- **C. Your classifier is generalizing poorly.**
- D. Your classifier is ready for use. **(No, will perform poorly on new data)**

Break & Quiz

- **Q 2.3**: You have trained a classifier, and you find there is significantly **lower** loss on the test set than the training set. What is likely the case?

- A. You have accidentally trained your classifier on the test set.
- B. Your classifier is generalizing well.
- C. Your classifier is generalizing poorly.
- D. Your classifier needs further training.

Break & Quiz

- **Q 2.3**: You have trained a classifier, and you find there is significantly **lower** loss on the test set than the training set. What is likely the case?

- **A. You have accidentally trained your classifier on the test set.** (This is very likely, loss will usually be the lowest on the data set on which a model has been trained)
- B. Your classifier is generalizing well.
- C. Your classifier is generalizing poorly.
- D. Your classifier needs further training.

Quiz break

Q1-1: K-NN algorithms can be used for:

- A Only classification
- B Only regression
- C Both

Quiz break

Q1-1: K-NN algorithms can be used for:

- A Only classification
- B Only regression
- C Both

Quiz break

Q1-2: Which of the following distance measure do we use in case of categorical variables in k-NN?

- A Hamming distance
- B Euclidean distance
- C Manhattan distance

Quiz break

Q1-2: Which of the following distance measure do we use in case of categorical variables in k-NN?

- A Hamming distance
- B Euclidean distance
- C Manhattan distance

Quiz break

Q1-3: Consider binary classification in 2D where the intended label of a point x = (x1, x2) is positive if x1>x2 and negative otherwise. Let the training set be all points of the form x = [4a, 3b] where a,b are integers. Each training item has the correct label that follows the rule above. With a 1NN classifier (Euclidean distance), which ones of the following points are labeled positive? Multiple answers.

- [5.52, 2.41]
- [8.47, 5.84]
- [7,8.17]
- [6.7,8.88]

Quiz break

Q1-3: Consider binary classification in 2D where the intended label of a point x = (x1, x2) is positive if x1>x2 and negative otherwise. Let the training set be all points of the form x = [4a, 3b] where a,b are integers. Each training item has the correct label that follows the rule above. With a 1NN classifier (Euclidean distance), which ones of the following points are labeled positive? Multiple answers.

- [5.52, 2.41]
- [8.47, 5.84]
- [7,8.17]
- [6.7,8.88]

Nearest neighbors are
[4,3] => positive
[8,6] => positive
[8,9] => negative
[8,9] => negative
Individually.

Quiz break

Q2-2: True or False
Maximum likelihood estimation is the same regardless of whether we maximize the likelihood or log-likelihood function.

- A True
- B False

Quiz break

Q2-2: True or False

Maximum likelihood estimation is the same regardless of whether we maximize the likelihood or log-likelihood function.

- A True
- B False

Quiz break

Q2-3: Suppose the weights of randomly selected American female college students are normally distributed with unknown mean μ and standard deviation σ. A random sample of 10 American female college students yielded the following weights in pounds:
115 122 130 127 149 160 152 138 149 180.

Find a maximum likelihood estimate of μ.

- A 132.2
- B 142.2
- C 152.2
- D 162.2

Quiz break

Q2-3: Suppose the weights of randomly selected American female college students are normally distributed with unknown mean μ and standard deviation σ. A random sample of 10 American female college students yielded the following weights in pounds:
115 122 130 127 149 160 152 138 149 180.

Find a maximum likelihood estimate of μ.

- A 132.2
- B 142.2
- C 152.2
- D 162.2

Quiz break

Q3-1: Which of the following about Naive Bayes is incorrect?

- A Attributes can be nominal or numeric
- B Attributes are equally important
- C Attributes are statistically dependent of one another given the class value
- D Attributes are statistically independent of one another given the class value
- E All of above

Quiz break

Q3-1: Which of the following about Naive Bayes is incorrect?

- A Attributes can be nominal or numeric
- B Attributes are equally important
- C Attributes are statistically dependent of one another given the class value
- D Attributes are statistically independent of one another given the class value
- E All of above

Quiz break

Q3-2: Consider a classification problem with two binary features, $x_1, x_2 \in \{0,1\}$. Suppose $P(Y = y) = 1/32$, $P(x_1 = 1 | Y = y) = y/46$, $P(x_2 = 1 \mid Y = y) = y/62$. Which class will naive Bayes classifier produce on a test item with $x_1 = 1$ and $x_2 = 0$?

- A 16
- B 26
- C 31
- D 32

Quiz break

Q3-2: Consider a classification problem with two binary features, $x_1, x_2 \in \{0,1\}$. Suppose $P(Y = y) = 1/32$, $P(x_1 = 1 | Y = y) = y/46$, $P(x_2 = 1 | Y = y) = y/62$. Which class will naive Bayes classifier produce on a test item with $x_1 = 1$ and $x_2 = 0$?

- A 16
- B 26
- C 31
- D 32

Quiz break

Q3-3: Consider the following dataset showing the result whether a person has passed or failed the exam based on various factors. Suppose the factors are independent to each other.
We want to classify a new instance with Confident=Yes, Studied=Yes, and Sick=No.

Confident	Studied	Sick	Result
Yes	No	No	Fail
Yes	No	Yes	Pass
No	Yes	Yes	Fail
No	Yes	No	Pass
Yes	Yes	Yes	Pass

- A Pass
- B Fail

Quiz break

Q3-3: Consider the following dataset showing the result whether a person has passed or failed the exam based on various factors. Suppose the factors are independent to each other.
We want to classify a new instance with Confident=Yes, Studied=Yes, and Sick=No.

Confident	Studied	Sick	Result
Yes	No	No	Fail
Yes	No	Yes	Pass
No	Yes	Yes	Fail
No	Yes	No	Pass
Yes	Yes	Yes	Pass

- A Pass
- B Fail

Quiz Break

Consider the linear perceptron with x as the input. Which function can the linear perceptron compute?

(1) $y = \text{ax} + \mathrm{b}$

(2) $y = \text{ax}^2 + bx + c$

A. (1)

B. (2)

C. (1)(2)

D. None of the above

Quiz Break

Consider the linear perceptron with x as the input. Which function can the linear perceptron compute?

(1) $y = \mathrm{ax} + \mathrm{b}$

(2) $y = \mathrm{ax}^2 + bx + c$

A. (1)
B. (2)
C. (1)(2)
D. None of the above

Answer: A. All units in a linear perceptron are linear. Thus, the model can not present non-linear functions.

Quiz Break

Perceptron can be used for representing:

A. AND function
B. OR function
C. XOR function
D. Both AND and OR function

Quiz Break

Perceptron can be used for representing:

A. AND function
B. OR function
C. XOR function
D. Both AND and OR function

Quiz Break

Which one of the following is valid activation function

a)Step function

b) Sigmoid function

C) ReLU function

D) all of above

Quiz Break

Which one of the following is valid activation function

a)Step function

b) Sigmoid function

C) ReLU function

D) all of above

Quiz Break

Let $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Which of the following functions is NOT an element-wise operation that can be used as an activation function?

A $\mathrm{f}(x) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

B $\mathrm{f}(x) = \begin{bmatrix} \max(0, x_1) \\ \max(0, x_2) \end{bmatrix}$

C $\mathrm{f}(x) = \begin{bmatrix} \exp(x_1) \\ \exp(x_2) \end{bmatrix}$

D $\mathrm{f}(x) = \begin{bmatrix} \exp(x_1 + x_2) \\ \exp(x_2) \end{bmatrix}$

Quiz Break

Let $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Which of the following functions is NOT an element-wise operation that can be used as an activation function?

A $\mathrm{f}(x) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

B $\mathrm{f}(x) = \begin{bmatrix} \max(0, x_1) \\ \max(0, x_2) \end{bmatrix}$

C $\mathrm{f}(x) = \begin{bmatrix} \exp(x_1) \\ \exp(x_2) \end{bmatrix}$

D $\mathrm{f}(x) = \begin{bmatrix} \exp(x_1 + x_2) \\ \exp(x_2) \end{bmatrix}$

Quiz Break

Which output function is often used for multi-class classification tasks?

A Sigmoid function

B Rectified Linear Unit (ReLU)

C Softmax function

D Max function

Quiz Break

Which output function is often used for multi-class classification tasks?

A Sigmoid function

B Rectified Linear Unit (ReLU)

C Softmax function

D Max function

Quiz Break

Suppose you are given a 3-layer multilayer perceptron (2 hidden layers h1 and h2 and 1 output layer). All activation functions are sigmoids, and the output layer uses a softmax function. Suppose h1 has 1024 units and h2 has 512 units. Given a dataset with 2 input features and 3 unique class labels, how many learnable parameters does the perceptron have in total?

Quiz Break

Suppose you are given a 3-layer multilayer perceptron (2 hidden layers h1 and h2 and 1 output layer). All activation functions are sigmoids, and the output layer uses a softmax function. Suppose h1 has 1024 units and h2 has 512 units. Given a dataset with 2 input features and 3 unique class labels, how many learnable parameters does the perceptron have in total?

$1024 * 2 + 1024 + 512 * 1024 + 512 + 512 * 3 + 3 = 529411$

Quiz Break

Consider a three-layer network with **linear Perceptrons** for binary classification. The hidden layer has 3 neurons. Can the network represent a XOR problem?

a) Yes

b) No

Quiz Break

Consider a three-layer network with **linear Perceptrons** for binary classification. The hidden layer has 3 neurons. Can the network represent a XOR problem?

a) Yes

b) No

Solution:

A combination of linear Perceptrons is still a linear function.

Quiz Break

Gradient Descent in neural network training computes the ______ of a loss function with respect to the model ______ until convergence.

A gradients, parameters

B parameters, gradients

C loss, parameters

D parameters, loss

Quiz Break

Gradient Descent in neural network training computes the _______ of a loss function with respect to the model _______ until convergence.

A gradients, parameters

B parameters, gradients

C loss, parameters

D parameters, loss

Quiz Break

Suppose you are given a dataset with 1,000,000 images to train with. Which of the following methods is more desirable if training resources are limit but enough accuracy is needed?

A Gradient Descent

B Stochastic Gradient Descent

C Minibatch Stochastic Gradient Descent

D Computation Graph